OR 610 – Deep Learning

Student name: Tuan Nguyen

Progress Summary: Long-term Rainfall Prediction Using LSTM Deep Learning Model

**I. Project Overview**

The purpose of this project is to create a deep learning model using recurrent architecture, more specifically LSTM type models, that can make accurate long-term prediction of 30 days for future rainfall based on observations of 60 days prior. This project aims to achieve 2 main goals, which are:

1. To accurately predict future rainfall for 30-days period, which is considered a difficult task for physical simulation models.
2. To only use observed rainfall as input variables, and not rely on other predictors such as windspeed, sunlight hours, max/min temperature, …etc. This allows the model to be deployed to lesser developed region/countries where there is insufficient coverage of hydrometeorological stations

**II. Current Progress**

This progress summary report serves as the midway point in the overall project progress mentioned in the project proposal:

* Gather data from 1980 – 2022:

(Completed – selected station: New Orleans Airport)

* Clean the data for anomalies:

Completed – data obtained from NOAA contain no anomalies

* Initial Visualization:

Completed – several graphs were created from the data shows monthly trend and overall

* Normalize data:

Completed – Min-max normalization was selected for this problem. However, since the maximum observed value in the past (332mm) might be surpassed in the future, the maximum selected for normalization is 350mm, which is considered incredibly unlike to be observed for single day rainfall event.

* Select input – output sizes:

Completed – 90 days sliding window was created from the raw input, where the first 60 days will be selected as input, the last 30 days will be the output

* Reformat data to correct dimensions:

Completed – all sliding windows was converted to tensors – final dimensions are (number of windows, 90)

* Split data into train – test – validation sets:

Completed – Data from 1980 to 2014 is selected for training purposes, data from 2015 – 2022 is withheld for model evaluation and will not be used during training. The train data from 1980 – 2014 is further split into train – test – validation with ratio of 0.75 – 0.1 – 0.15 respectively, with train – test -validation set each contains 9522 – 1269 – 1905 windows. 2015 – 2022 data for final model evaluation contains 2833 windows.

* Select loss function:

Completed – 2 possible loss functions that is suitable for regression problem that is available on PyTorch are: torch.nn.L1Loss (Mean Absolute Error), and torch.nn.MSE (Mean Squared Error). In this project, L1 loss is preferred due to its ability to handle outliers. Histogram of the data shows that the majority of days have <3mm rainfall (insignificant rainfall), which make rainfall events outliers in general.)

* *Project summary (Apr 8)*
* Create LSTM model architecture (10 days – until Apr 18):

In Progress – Limit placed on the input data means that the model architecture needs to have sufficient ability to make accurate predictions with limited data and without overfitting.

* Complete Code for model training and evaluation (parallel with above task):

In Progress – The code for training will follow similar patterns/method with provided RNN examples and HW3

* Grid search for suitable hyperparameters (4 - 5 days – until Apr 23):

In queue – after finalizing model architecture and training code

* Finalize the model and measure performance. (2-4 days – until Apr 26):

In queue – after finalizing model architecture and parameters

* Visualize result and compare predictions (2-4 days – until Apr 30):

In queue – after finalizing model architecture and parameters

* Complete and submit the report (parallel with all above tasks)

**III. Risk and Issues**

1. Model complexity

One of the goals of this project is to be able to produce long-term accurate predictions using only rainfall input. As the input is restrained, the model will need to have sufficient complexity to capture the relationship between rainfall from previous 60 days and 30 days in the future, but not too complex as to avoid overfitting and costly training. A possible mitigation plan is to do further research on effective/popular LSTM architectures, references similar model type used in different field such as long-term stock prediction, and/or increase the input window to 90 – 150 days.

1. Insufficient input data dimensions

Rainfall is part of the complex global hydrological cycle. Thus, it is influenced by a variety of factors, such as temperature, air pressure, wind speed, ocean currents, and even climate change. While this project aims to lower the input requirements of a long-term prediction model, there are risks of valuable information existing within measurement of other factors not being exploited to increase prediction performance. This risk is mitigated by including other available factors such as wind speed, temperatures, sunlight hours in the model input. In case that the New Orleans Airport stations have insufficient data related to these factors, data from other stations can be selected as model input instead.

1. Non-deterministic predictions

This occurs when uncertainty/inaccurate prediction at a timestep can eventually lead to predictions that do not closely reflect reality, due to the uncertainty is amplified through each timestep. This butterfly effect is also observed in physical simulation models and is considered one of the reasons why long-term prediction is challenging. In this project, the mitigation plan is to have the model predict 30 days simultaneously, and not to predict 1 time step at a time.

**IV. Conclusion**

Overall, the project is on a steady pace to completion by the time of the presentation. The project plan has details on tasks and their requirements, and appropriate mitigation plans are available for possible project risk and issue. Other in-progress tasks and issues will be addressed in the next coming weeks.

**V. Appendix**

Contains visualizations about the data

Chart, histogram

Description automatically generated

Figure 1: Histogram of rainfall value (mm)

Chart

Description automatically generated

Figure 2: Frequency of daily rainfall by month

Calendar

Description automatically generated with medium confidence

Figure 3: Daily Average rainfall for significant rainfall events (>3mm)